CHARACTERIZATION OF ERRORS IN TRANSCRIPTOMICS BY SPOTTED DNA MICROARRAY DATA

Tobias K. Karakach and Peter D. Wentzell,
karakach@dal.ca and peter.wentzell@dal.ca,
Department of Chemistry,
Dalhousie University

ABSTRACT

The so-called '*-omics* revolution' is characterised by high throughput measurements and vast quantities of data. Unfortunately, the quality of the measurements is often over-estimated and, as such, simplistic assumptions regarding the structure of the measurement errors are made subsequent to application of routine data analysis techniques. In transcriptomics by spotted DNA microarrays, it is beginning to emerge that the conceptual simplicity of the technology belies a potentially complex measurement error structure. For example it has been shown variously that the data exhibit a non-constant variance leading to the recent development of models to 'fix' the problem. Regrettably, the extent to which these error inhomogeneities affect individual microarray data has not been systematically explored and, consequently, the structure of the measurement uncertainties after transformations are applied remains uncharacterized.

This work presents the results of a systematic characterization of the measurement error structure for spotted DNA microarray data as well as a model that compartmentalizes the total variance exhibited by the measured intensity ratios into distinct components that can be measured independently. In particular, one of the ratio variance components, which is often ignored in microarray data analysis, is the uncertainty associated with the measurement of the ratio itself. In most microarray data, the ratio is determined as a mean or median of pixel intensities comprising a spot and no implicit information is provided about the accuracy with which this quantity is measured. In this work, the ratio is measured as an orthogonal slope of the pixel intensities comprising the spot and a bootstrap approach is employed in determining the magnitude of the uncertainty associated with determining this ratio. Those measurements for which this value dominates the total variance are eliminated in order to maintain the distribution of errors. The structure of the measurement uncertainties is then characterized empirically using replicate measurements.

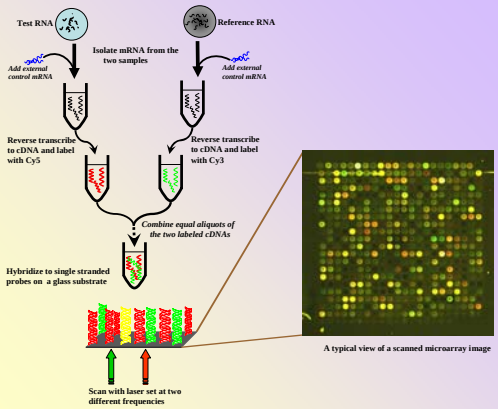
INTRODUCTION

DNA microarrays are an emerging technology for genome-wide analysis of gene expression. Microarrays could potentially be used for investigating the onset and prognosis of diseases as well as investigating the effect, to cells, of drugs by observing the global over- or under-expression of gene products under these conditions. The heart of this technology lies in the hybridization of complementary DNA sequences.

Single stranded gene sequences representing an organism are immobilized on a glass chip and used to probe for differential concentration of complementary pairs of similar genes derived from the organism – under different conditions. Although conceptually simple, the complexities of the experimental processes involved in this technology often introduce random and systematic biases into the measurements. These biases could be large enough to invalidate the effects that are under investigation. Among other sources of variability, systematic biases in microarray data can be attributed to the differential concentration and amount of cDNA placed on the microarray slides, spotting pins that may wear out over time, hybridization efficiency, mRNA preparation, lack of spatial homogeneity of the hybridization on a slide, dye biases and scanner settings. Random variability results in spot images that can show characteristics of poor definition, unusual morphology, low intensity, high background and signal saturation among other features.

Although some of the variability can be controlled by the experimenter to a limited extent, few can be completely eliminated. It is therefore prudent to remove the effects of such systematic variations and bring the data, collected on different scales onto a common one.

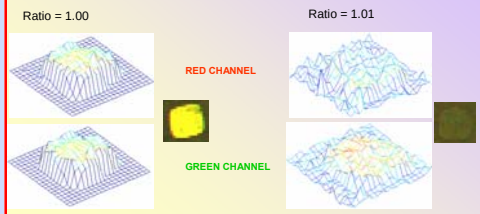
THE GENERAL EXPERIMENTAL SETUP



- The general experimental setup of the DNA microarrays used for gene expression analysis is as shown in this Figure
- From this setup, it is possible to envisage several stages at which uncertainty may be introduced into the experiments

DATA QUALITY ANALYSIS

- The analysis of microarray data typically uses ratios or $\log_2$ ratios in order to account for within spot and between spot variations
- Unfortunately, most of the ratio measurements do not carry implicit or explicit information about the inherent uncertainty. As exemplified below, measurements with the same ratio can have radically different physical characteristics



DATA QUALITY ANALYSIS

- Over time, there has been a recognition that microarray data are inherently noisy and subsequent remedies have been devised to mitigate the noise
- These approaches can be categorized as "data filtering methods" and "measurement error models"
- Data filtering methods strive at identifying and excluding spot images, from further analysis, based on visual inspection to identify technical flaws
- Some of the disadvantages of this approach include their labour intensive nature as well as their subjectivity to operator bias and image representation by the software
- In addition, the binary classification of spots as good or bad assumes that there is no continuum in the quality of spots
- Further, exclusion of spots by data filtering methods, requires that the spot is classified as missing in subsequent arrays in the analysis of multiple arrays

DATA QUALITY ANALYSIS

- Measurement error models for microarrays are based on some statistical distributional assumptions and emphasize reclaiming an assumed error structure
- Some of these methods are overly complex requiring either replicate data (which is seldom available) or estimation of several parameters [1 - 3]

OBJECTIVES OF THE CURRENT WORK

- To develop a method for estimating the variance in the ratio measurement, based solely on the morphological characteristics of the spot image and without the need for replicate data
- To characterize measurement error variance in spotted DNA microarray data in view of the anticipated proportional error structure and additive components of the errors

PROPOSED ERROR MODEL

- The model proposed is a nested summary of the variability in spotted DNA microarray data

$$\sigma^2_{rat} = \sigma^2_{biol} + \sigma^2_{slide} + \sigma^2_{spot} + \sigma^2_{meas}$$

Biological variation | Spatial variation | Variation between spots | Uncertainty in the measurement process

- This model can be simplified as

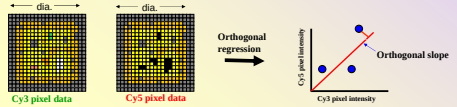
$$\sigma^2_{rat} = \sigma^2_{expt} + \sigma^2_{meas}$$

Proportional error component | Additive error component

where the first term on the right hand side depicts a proportional component of the total variance in the ratio, and can be estimated using replicates, while the last term is the additive component that must be determined for each spot

PROPOSED ERROR MODEL AND RATIO CALCULATION

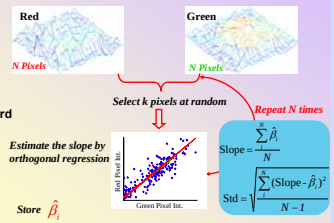
- Estimating the last term in this model - the variance associated with the evaluation of the ratio from pixel data - is the focus of the first objective of this work. This will be influenced by a variety of factors, including spot morphology, signal intensity and background measurements
- It is important to identify those cases where σ^2_{meas} dominates the overall variance so that these measurements can be excluded or weighted appropriately
- In order to achieve these objectives, the appropriate ratio calculation method was identified as the orthogonal regression of the pixel intensities on the two channels
- This approach for ratio calculation eliminates the need for background estimation given that the intercept mitigates the differential background between the two channels
- Procedurally, select red/green (x/y) pairs for the regression, excluding: (a) pixels at saturation, (b) five most intense red pixels and, (c) five most intense green pixels



ESTIMATING MEASUREMENT UNCERTAINTY

Bootstrapping Method

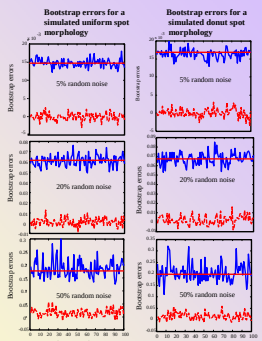
- The uncertainty in the regression ratio was estimated through bootstrapping
- This is done by randomly re-sampled (with replacement) the pixel data *N* times, repeating the calculation each time
- The ratio was reported as the mean of the orthogonal bootstrap slopes
- The uncertainty in the ratio was reported as the standard deviation of the bootstrap slopes



ESTIMATING MEASUREMENT UNCERTAINTY

Validating the Bootstrap Estimated Errors

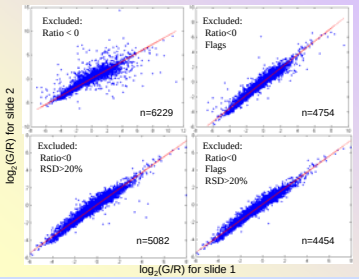
- 100 microarray spots that mimicked two typical morphologies were simulated and bootstrap estimates of the uncertainty determined
- Different levels of random noise realizations were added to the spots prior to the bootstrap estimates
- The standard deviation of the ratios from the 100 spots were also obtained and the agreement between this standard deviation and the bootstrap estimates is shown
- Results in this figure indicate that at least the bootstrap estimation method is valid for ideal, simulated data
- Experimental validation is difficult since it requires multiple replicate spots on the array, which is not often available for microarrays
- Bootstrap error estimates were therefore assumed to be representative
- Subsequently, those cases where the bootstrap error dominated the total variance were identified and eliminated. Such cases were hypothesised to destroy the proportional error structure for microarrays
- This hypothesis was tested by eliminating measurements whose bootstrap RSD was greater than a given threshold, and examining the error distribution of the remaining data



* Note: the red dotted line indicates the bias in bootstrap estimates of the ratio, the red straight line is the standard error of the 100 ratios while the blue line is the bootstrap errors

CHARACTERIZING MEASUREMENT UNCERTAINTY

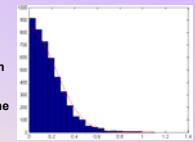
- Ideally, log-ratio vs. log-ratio plots of replicates should be uniformly distributed around the regression line (the errors should exhibit a uniform distribution) and exhibit unity slope and a zero intercept
- In practice, this is observed only when measurements with high bootstrap RSD are eliminated as shown in the figures here

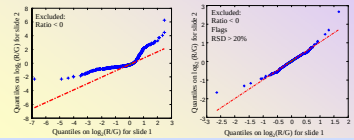


CHARACTERIZING MEASUREMENT UNCERTAINTY

Distribution of Orthogonal Residuals

- The distribution of orthogonal residuals for the preceding plot (censored for RSD> 20%, ratio < 0 and flags) appears normal as shown in this histogram
- This suggests that the overall error in the ratio, corrected for σ^2_{meas}, is proportional
- Q-Q plots for the *x-y* orthogonal residuals before and after censoring show a restoration of a normal distribution residuals





CONCLUSIONS

- In order to fully exploit DNA microarray data, much more attention must be paid to the quality of the ratio data provided
- The nested measurement error model proposed, partitions the uncertainty in the ratio into a constant proportional component and an additive component, which is the uncertainty in the ratio measurement process
- The relative uncertainty in the ratio measurement step is highly variable and needs to be taken into account at higher levels of data analysis
- Bootstrap methods proved to be fairly reliable for estimating the uncertainty in the ratio. Screening points on the basis of the relative standard deviation (coefficient of variation) in these ratios was effective for removing unreliable measurements
- When points with high measurement errors are removed, the residuals of the log-log plots for replicates appear to follow a normal distribution suggesting that the other component, σ^2_{expt} of the uncertainty in the ratio is a proportional error contribution

References:

1. Rocke, D. M. and Durbin, B.; *J. Comput. Biol.*, (2001), **8**: 557-569
2. Ideker, T., Thorsson, V., Seigel, A. F., and Hood, L. E.; *J. Comput. Biol.*, (2000), **7**: 805 – 817
3. Goryachev, A. B., Macgregor, P.F., and Edwards, A.M.; J. Comput. Biol., (2001), **8**: 443-461

Acknowledgements:

Natural Sciences and Engineering Research Council of Canada

S. Douglas, V. Ewart and J. Williams @ National Research Council Of Canada